Investigación Educativa
vol. 13 N.º 24, 135-155
Julio-Diciembre 2009,
ISSN 1728-5852

INFORMACIÓN ADICIONAL OBTENIDA CON EL MODELO RASCH

ADDITIONAL INFORMATION OBTAINED WITH THE RASCH MODEL

Fecha de recepción: 14/10/2009 Fecha de aceptación: 05/12/2009

Luis Hurtado Mondoñedo[1]

RESUMEN

La evaluación de los aprendizajes comprende el recojo de información, análisis de esta información, juicio sobre el resultado de este análisis y la toma de decisiones de acuerdo con el juicio emitido. Recoger la información se convierte en la primera etapa de este proceso y en ella juegan un importante papel las prácticas calificadas y exámenes. Ellos son los instrumentos con los que recogemos información que, después de ser analizada, respaldan los informes acerca del aprendizaje de los alumnos. Este análisis puede hacerse en distintos grados. La teoría moderna proporciona un marco teórico que permite interpretar de manera más fina los puntajes observados. Aquí se incluye el modelo de Rasch. El objetivo del presente artículo es mostrar la información adicional que se puede obtener al analizar una prueba siguiendo el modelo de Rasch.

Palabras clave: medida ordinal, escala de intervalo, modelo de Rasch, confiabilidad, medidas de ajuste.

1 Magíster en educación en la mención de Medición, Evaluación y Acreditación de la Calidad de la Educación por la Universidad Nacional Mayor de San Marcos. Investiga acerca de Medición educacional siguiendo la Teoría Clásica y Moderna del Test.
E-mail: luchohurtado@yahoo.com.

Art. Revista Investigacion Educativa N 24.indd 135 24/05/2010 03:30:06 p.m.

ABSTRACT

Learning evaluation includes the gathering of information, analysis of the information, interpretation of the obtained results and decision- making according to the analysis performed. The gathering of information then becomes the first stage of this process and the role of graded quizzes and exams are very important. These are the assessment tools with which we gather information that, after being analyzed, back up the reports about the student's learning process. This analysis can be made in different degrees. Modern Theory provides the theoretical frame which allows a more refined interpretation of the observed grades. Here the Rasch's Model is important. The purpose of the present article is to show additional information that can be gathered when analyzing student's grades.

Key words: ordinal measurement, interval scale, Rasch's model, reliability, fit's measurements.

INTRODUCCIÓN

Lord Kelvin, citado por Sears y Zemansky (1981), señalaba que "… nuestro saber es deficiente e insatisfactorio mientras no somos capaces de expresarlo en números" (p.3). Los científicos físicos lograron grandes avances debido a que los conceptos estudiados eran factibles de ser medidos. La medición por mucho tiempo estuvo asociada con los fenómenos físicos. En la década de 1930 se discutió acerca de si era posible la estimación cuantitativa de los eventos sensoriales. En aquella época las medidas tomadas por los científicos sociales no cumplían las reglas lógicas señaladas por la física, entre ellas la operación de adición se convertía en el mayor obstáculo de la medición. Los científicos sociales dirigidos por Stevens enfrentaron el asunto definiendo un nuevo tipo de medición: "Medición es la asignación de números a objetos o eventos de acuerdo a una regla". Esta definición deja de lado la noción de magnitud e incorpora el concepto de "nivel de medición" lo que permitió el desarrollo de las escalas de medición en los niveles ordinal, nominal, intervalar y de ratio. Los puntajes de los alumnos en una prueba, obtenido por la suma aritmética de las puntuaciones de las preguntas, es un ejemplo de medición en el nivel ordinal. A este nivel tenemos argumentos muy débiles para hacer inferencias acerca de los aprendizajes de los alumnos.

Art. Revista Investigacion Educativa N 24.indd 136 24/05/2010 03:30:06 p.m.

1. MARCO TEÓRICO

1.1. Los puntajes de la prueba: una medida ordinal

Los puntajes de los alumnos se convierten en un indicador del aprendizaje de los contenidos incluidos en la prueba. Sin embargo el puntaje del alumno no mide realmente el nivel de dominio del tema. Este solo ubica al alumno dentro de una escala numérica ordinal. En una escala ordinal usamos la propiedad del orden. Si el número que corresponde al puntaje de un alumno P es mayor que el número que corresponde al puntaje de otro alumno Q, entonces decimos que P posee mayor conocimiento en el tema en cuestión que Q. Al ordenar a los alumnos según su puntaje es posible detectar diferencias de grado en los niveles de dominio del tema aunque de forma imprecisa. Un concepto importante en la medición es el de linealidad. Cuando utilizamos una cinta métrica, una balanza o un termómetro estamos frente a medidas lineales. Esto significa que una unidad más en la escala significa siempre una unidad más. Así por ejemplo, la diferencia entre 37 Kg y 36 Kg es la misma que entre 73 Kg y 72 Kg. La distancia entre ellas significa lo mismo a lo largo de la escala. Esta propiedad no se cumple en medidas ordinales como lo son los puntajes brutos. De acuerdo a Glass y Stanley (1970) en una medición ordinal "los resultados de la aritmética no pueden ser interpretados como diciendo algo acerca de las cantidades en la propiedad poseída actualmente por los objetos" (p.10). Diferencias iguales en los puntajes no implica igual diferencia en el dominio de los contenidos tratados en la prueba. Si tenemos tres alumnos Jenny, Sara y Mario con puntajes de 15, 12 y 9 puntos respectivamente, la diferencia entre los puntajes de Jenny y Sara es igual que la diferencia entre los puntajes de Sara y Mario. Pero esto no significa que la diferencia relativa entre sus conocimientos sea la misma. Lo "más" que sabe Jenny con respecto a Sara no es igual a lo "más" que sabe Sara con respecto a Mario.

1.2. El modelo de Rasch y la escala de intervalo

Una escala en la que la interpretación de las diferencias es igual a lo largo de ella se dice que posee la propiedad de intervalo. Las medidas de longitud, peso o temperatura forman una escala de intervalo. Esto permite lograr una medida más precisa debido a que los instrumentos usados para su medición se encuentran debidamente calibrados. En una escala de intervalo se hace posible que una unidad más sea siempre una unidad

Art. Revista Investigacion Educativa N 24.indd 137 24/05/2010 03:30:06 p.m.

más. Las pruebas de rendimiento también constituyen instrumentos de medición. Las medidas estimadas a partir de ellas deberían formar una escala de intervalo. La familia de los modelos de Rasch hace posible esta escala logrando con ello una medición más fina. El más simple de estos es el modelo dicotómico – acierto y falla – el cual, aplicado a las pruebas de rendimiento, toma en cuenta tanto el nivel de dominio de los alumnos como la dificultad de las preguntas contenidas en la prueba. Este postula que la probabilidad "P_{si}" que tiene un alumno "s" de responder correctamente una pregunta "i" depende de la diferencia entre el nivel de dominio del alumno "B_s" y el nivel de dificultad de la pregunta "D_i" cuya formulación matemática está dada por:

$$P_{si} = \frac{e^{Bs - Di}}{1 - e^{Bs - Di}} \qquad \ldots (I)$$

1.3. La escala LOGIT

Para Thurstone toda medición es siempre una abstracción y la medición de un objeto es ubicar el objeto en un punto de un continuo abstracto (Wright y Stone, 1999). Bajo el modelo de Rasch los niveles de habilidad de los alumnos y dificultad de las preguntas se ubican en la misma escala y presentan las mismas unidades. Las unidades son los "logits" por lo que nombraremos la escala como LOGIT. De la ecuación I podemos obtener la siguiente expresión equivalente:

$$1n \left(\frac{P_{si}}{1 - P_{si}} \right) = B_s - D_i$$

Un ratio de probabilidades es llamado "odds". El primer miembro de la igualdad nos presenta el logaritmo natural de este ratio. De esta forma la diferencia $B_s - D_i$, y por tanto B_s y D_i, están medidas en unidades del logaritmo del ratio "log-odds units", de ahí el nombre "logits". La escala LOGIT cubre todos los números reales aunque por lo general las medidas estimadas se encuentran entre -5.0 y 5.0 logits. Tristán (2001) señala que "esta medida tiene la ventaja de que limita los valores de medida a intervalos razonables y permite tomar en cuenta tanto el éxito como el fracaso en una sola cantidad" (p.11). En una prueba calificada con una convencional escala ordinal de 0 a 20 puntos, el puntaje mínimo de 0 puntos no significa que el alumno no sabe nada del tema, del mismo

Art. Revista Investigacion Educativa N 24.indd 138 24/05/2010 03:30:06 p.m.

modo el puntaje máximo de 20 puntos no se interpreta como que el alumno sabe todo del tema. Esto solo permite ubicar a los alumnos en orden decreciente de acuerdo al puntaje bruto obtenido. Las inferencias hechas a partir de esta medida ordinal son débiles. De acuerdo con Wright y Mok (2004) a fin de construir la inferencia de la observación, el modelo de medición debe producir medidas lineales, superar los datos que faltan, proporcionar estimaciones de la precisión, contar con dispositivos para la detección de desajuste, además los parámetros del objeto que se mide y del instrumento de medición deben ser separables. La familia del modelo de Rasch permite resolver este problema.

1.4. Estimaciones de la precisión

La confiabilidad está relacionada con la precisión del instrumento. Los resultados obtenidos de la prueba serán más confiables cuanto menor sea el error contenido en ella. En el modelo de Rasch el error está asociado con la diferencia entre la respuesta observada a una pregunta y la probabilidad de responderla correctamente. El tradicional coeficiente de confiabilidad de la prueba (ρ) definido por Spearman como la razón de las puntuaciones verdaderas y las puntuaciones observadas es estimado por el modelo de Rasch a partir de dos coeficientes. Para los alumnos el REAL RELIABILITY y el MODEL RELIABILITY proporcionan respectivamente una cota inferior y superior de ρ. El coeficiente de confiabilidad varía entre 0 y 1, cuanto más cercano a 1, menor error de medición por tanto mayor precisión en la estimación de las medidas. No existe un estándar absoluto que pueda ser usado para señalar si un coeficiente de confiabilidad es lo suficientemente alto. Esto depende de las decisiones que se vayan a tomar a partir de los puntajes observados en la prueba. Frisbie (1988) señala que no hay un estándar para juzgar si un coeficiente de confiabilidad es alto y "usualmente podemos tolerar coeficientes alrededor de 0.50 para puntajes en pruebas hechas por docentes de aula si cada puntaje va a ser combinado con otro tipo de información" (p.29). Para las preguntas el ITEM RELIABILITY no tiene un significado equivalente a ρ. Indica la reproducibilidad de la ubicación de las preguntas a lo largo del continuo, si estas mismas fueran aplicadas a otra muestra similar (Bond y Fox, 2007). Una confiabilidad alta significa que hay una alta probabilidad que las preguntas estimadas con altas medidas tengan medidas más altas (dificultad) que las preguntas con bajas medidas en una segunda aplicación de la prueba.

Art. Revista Investigacion Educativa N 24.indd 139 24/05/2010 03:30:07 p.m.

1.5. Ajuste de los datos al modelo

Siguiendo un patrón lógico, un alumno debería responder correctamente aquellas preguntas cuyos niveles de dificultad sean menores que su nivel de habilidad. No podría responder correctamente aquellas con niveles de dificultad mayores que su nivel de habilidad. Esto es lo lógico y por tanto lo más probable sin embargo al analizar el patrón de respuestas de los alumnos se puede notar que no siempre ocurre de ese modo. Existen desajustes en los datos recogidos. El modelo de Rasch nos proporciona dos medidas de ajuste y pueden realizarse tanto para los alumnos como para las preguntas. El ajuste interno o INFIT mide el comportamiento inesperado del alumno (pregunta) en las preguntas (alumnos) cercanas(os) a su nivel de habilidad (dificultad). Un alumno que falla muchas preguntas cercanas a su nivel de habilidad tendría un alto INFIT, una pregunta que no es respondida correctamente por alumnos cercanos a su nivel de dificultad, también tendría un alto INFIT. El ajuste externo u OUTFIT mide el comportamiento inesperado del alumno (pregunta) en las preguntas (alumnos) alejadas(os) de su nivel de habilidad (dificultad). Un alumno que acierta muchas preguntas alejadas de su nivel de habilidad tendría un alto OUTFIT, una pregunta que es respondida correctamente por alumnos alejados de su nivel de dificultad, también tendría un alto OUTFIT. Valores MNSQ para las medidas de ajuste en el rango de 0.8 a 1.2 muestran un comportamiento razonable (Bond y Fox, 2007) siendo productivos para la medición los comprendidos entre 0.5 y 1.5 (Linacre, 2007). Así mismo valores ZSTD entre -2 y +2 indican un ajuste razonable (Tristán, 2001; Gonzáles, 2008).

1.6. Independencia de los parámetros

Uno de los aportes de la teoría moderna es que posibilita obtener medidas invariantes respecto de los sujetos involucrados e instrumentos utilizados (Muñiz, 1997; Hambleton, Swaminathan y Rogers, 1991; Andrich, 1988). Este es uno de los fundamentos de la medición. Los objetos medidos deben ser independientes del instrumento que utilicemos para medirlo, así mismo, el instrumento de medida debe ser independiente de los objetos medidos. En cuanto a pruebas de rendimiento esto significa que las medidas de los alumnos deben ser independientes de las preguntas con las que ellas fueron estimadas, y que las dificultades de las preguntas no dependen de las características del grupo sobre las que

Art. Revista Investigacion Educativa N 24.indd 140 24/05/2010 03:30:07 p.m.

fueron estimadas. A diferencia de la teoría clásica los modelos de la teoría moderna nos permiten obtener medidas independientes. Bajo el enfoque clásico una pregunta resulta fácil o difícil según la aptitud de los alumnos que fueron examinados, y la aptitud de los examinados depende de si las preguntas de la prueba fueron fáciles o difíciles. Este problema conceptual hace muy difícil comparar alumnos que rindieron distintas pruebas, y muy difícil comparar preguntas cuyas características son obtenidas usando diferentes grupos de alumnos (Hambleton, Swaminathan y Rogers, 1991). El modelo de Rasch tiene en cuenta un solo parámetro, el de dificultad de las preguntas, el cual está en las mismas unidades que las medidas de los alumnos. Contar con un banco de preguntas, donde se conocen sus parámetros, y aplicar el modelo de Rasch hace posible estimar el nivel de dominio del alumno sin caer en las debilidades de la teoría clásica.

2. MÉTODO

2.1. Participantes e Instrumento

Los datos corresponden a un grupo de 75 alumnos del grupo de estudios LA MATRIZ inscritos en el ciclo de preparación para rendir el examen de admisión a la Universidad del Pacífico (UP) durante el verano del 2009. La prueba analizada correspondió al bloque de preguntas de Aritmética del primer simulacro de admisión aplicado a los alumnos. Las preguntas fueron construidas tomando como base los temas del cuestionario de admisión UP 2009 los mismos que fueron desarrollados durante el ciclo de preparación. La prueba comprendió 10 preguntas, cada una presentaba un enunciado y cinco opciones de las cuales solo una era correcta.

3. Resultados y discusión

Las respuestas de los alumnos a las preguntas fueron codificadas dicotómicamente, con 1 para la respuesta correcta y 0 para la incorrecta. La base de datos así formada fue analizada con el programa SPSS 12 y Ministep (Linacre, 2007). El análisis de la prueba se inicia presentando el resumen de la información para los alumnos y las preguntas. La tabla I muestra la información básica de los alumnos. Esta presenta los resultados de 73 alumnos (NON-EXTREME), no incluyendo los 2 alumnos que obtuvieron el mínimo puntaje. En la columna de los puntajes (RAW SCORE) se muestra la media de 4.5 y la D.S. de 2.3. El puntaje máximo es de 9 puntos y el mínimo de 1 punto. En la columna de las medidas

Art. Revista Investigacion Educativa N 24.indd 141 24/05/2010 03:30:07 p.m.

(MEASURE) notamos que los máximos y mínimos fueron estimados en 2.42 y -2.51 logits respectivamente, con una media de -0.30 logits y una D.S. de 1.29 logits. La columna del ajuste interno (INFIT) reporta valores para el MNSQ comprendidos entre 0.52 y 1.62, con una media de 1.01 indicando que existe un buen ajuste interno. Para el ajuste externo (OUTFIT) los valores del MNSQ se encuentran comprendidos entre 0.22 y 3.28, con una media de 0.99. De acuerdo a la media podemos decir que, en general, existe un buen ajuste externo; sin embargo, el valor máximo 3.28 es un indicador de cierto ruido que debemos analizar.

Tabla I

SUMMARY OF 73 MEASURED (NON-EXTREME) alumnos

	RAW			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	4.5	10.0	-.30	.79	1.01	.1	.99	.1
S.D.	2.3	.0	1.29	.14	.28	.9	.48	.8
MAX.	9.0	10.0	2.42	1.12	1.62	2.7	3.28	2.9
MIN.	1.0	10.0	-2.51	.68	.52	-1.9	.22	-1.4
REAL RMSE	.86	ADJ.SD	.96	SEPARATION	1.12	alumno RELIABILITY		.56
MODEL RMSE	.81	ADJ.SD	1.01	SEPARATION	1.25	alumno RELIABILITY		.61
S.E. OF alumno MEAN = .15								

MINIMUM EXTREME SCORE: 2 alumnos

La tabla II muestra la información básica de las 10 preguntas de la prueba. En este caso no hubo preguntas extremas, todas fueron contestadas. La columna RAW SCORE muestra el número de respuestas correctas de las preguntas. La media fue de 32.6 y la D.S. de 10.5. El valor máximo fue de 58 y el mínimo de 20 respuestas correctas. En la columna de las medidas notamos que los máximos y mínimos fueron estimados en 1.01 y -2.09 logits respectivamente, con una media de 0.00 logits y una D.S. de 0.85 logits. La columna del INFIT reporta valores para el MNSQ comprendidos entre 0.80 y 1.25, con una media de 1.00. Para el ajuste externo (OUTFIT) los valores del MNSQ se encuentran comprendidos entre 0.70 y 1.57, con una media de 0.99. Los valores anteriores nos indican que las preguntas de la prueba de aritmética presentan un buen ajuste interno y externo.

Art. Revista Investigacion Educativa N 24.indd 142 24/05/2010 03:30:07 p.m.

Tabla II

SUMMARY OF 10 MEASURED (NON-EXTREME) preguntas

	RAW			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	32.6	73.0	.00	.28	1.00	.0	.99	.0
S.D.	10.5	.0	.85	.02	.12	1.0	.22	.9
MAX.	58.0	73.0	1.01	.33	1.25	1.9	1.57	2.3
MIN.	20.0	73.0	-2.09	.27	.80	-1.7	.70	-1.2
REAL RMSE	.29	ADJ.SD	.80	SEPARATION	2.73	pregun RELIABILITY		.88
MODEL RMSE	.28	ADJ.SD	.80	SEPARATION	2.80	pregun RELIABILITY		.89
S.E. OF pregunta MEAN = .28								

3.1. Reporte de alumnos y preguntas

La tabla III muestra el reporte general para el total de alumnos (75) que rindieron la prueba de aritmética. La primera columna (ENTRY NUMBER) indica el lugar que ocupa el patrón de respuestas de cada alumno en la data original. La segunda columna (TOTAL SCORE) muestra el número de respuestas correctas del alumno en la prueba. La columna COUNT indica el número de puntos utilizados para la construcción de las medidas. La columna MEASURE muestra las medidas estimadas de los alumnos. Podemos notar que los alumnos con el mismo score presentan la misma medida aunque no siempre el mismo ajuste. Según el modelo de Rasch, si bien el puntaje total contiene toda la información necesaria para la estimación de la capacidad de la persona, esto no es suficiente para ver si las respuestas observadas encajan en el modelo.

En general los resultados indicaron que los alumnos presentan un buen ajuste interno ya que en el 97% de ellos se observa un INFIT dentro del intervalo]0.5, 1.5[. En cuanto al ajuste externo encontramos que 5 alumnos presentan un OUTFIT inferior a 0.5 y 8 un valor mayor que 1.5, aunque en solo uno de ellos (A48) se observa un valor crítico. Este corresponde al máximo del OUTFIT y que anteriormente fue advertido como indicador de ruido. Una de las ventajas que otorga el análisis con el Modelo de Rasch es que este nos permite detectar algunos desajustes

Art. Revista Investigacion Educativa N 24.indd 143 24/05/2010 03:30:07 p.m.

en las respuestas dadas por los alumnos. El OUTFIT de 3.28 estaría indicando un comportamiento inesperado en las preguntas alejadas del nivel de habilidad del alumno A48. Para una mejor comprensión en la tabla IV presentamos una parte del escalograma de Guttman para las respuestas dadas por los alumnos, resaltando el patrón de respuestas del alumno A48. En el escalograma las preguntas están ordenadas de menor a mayor dificultad, y podemos notar que este alumno respondió incorrectamente las preguntas P3 y P5 cuyas medidas estimadas en logits (D_3=-2.09 y D_5=-0.54) se encuentran bastante por debajo del nivel estimado del alumno (B_{15}=0.99 logits). Aquí también se puede notar que este alumno respondió incorrectamente la pregunta P4 (D_4=0.13 logits) lo que se traduce en un INFIT de 1.59 indicador de un comportamiento inesperado en una pregunta cercana a su nivel de habilidad.

La tabla V muestra el reporte general para las 10 preguntas de la prueba de aritmética. Las columnas tienen los mismos significados que los descritos para la tabla III. En general los resultados indicaron que las preguntas presentaron un buen ajuste interno y externo. En todas se observó un INFIT dentro del intervalo productivo para la medición y solo una de ellas P1 mostró un OUTFIT (1.57) muy próximo fuera de él.

El modelo de Rasch nos permite hallar la probabilidad que tiene un alumno con un nivel de habilidad (B) de responder correctamente una pregunta con un nivel de dificultad dado (D). Si B>D lo más probable es que el alumno responda correctamente la pregunta; si B<D lo más probable es que el alumno responda incorrectamente - o no responda - la pregunta. Si B=D la probabilidad de éxito es 0.5 y nosotros podríamos esperar igualmente una respuesta correcta o incorrecta. Si bien esto es lo más lógico, no siempre ocurre de ese modo. Pueden encontrarse desajustes en algunas preguntas con respecto al modelo de medición empleado. Estos se pueden detectar a partir del OUTFIT.

Con ayuda de la tabla VI podemos analizar el desajuste mostrado por la pregunta P1. En las filas se muestran las preguntas ordenadas según el OUTFIT MNSQ y en las columnas los alumnos ordenados de izquierda a derecha de mayor a menor nivel de habilidad. El desajuste de P1 es debido a que cuatro alumnos (A54, A42, A72 y A29) con niveles de habilidad menores al nivel de dificultad de P1 la respondieron correctamente. La tabla V indica una medida de D_1=0.43 para P1 mientras que las medidas de los alumnos fueron estimadas en B_{54}=-1.57, B_{42}=-1.57,

Art. Revista Investigacion Educativa N 24.indd 144 24/05/2010 03:30:07 p.m.

Tabla III

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	alumno
19	9	10	2.42	1.07	1.09	.4	1.10	.5	.05	90.0	90.0	A19
3	8	10	1.57	.82	.80	-.3	.60	-.3	.49	80.0	80.0	A03
4	8	10	1.57	.82	.91	-.1	.71	-.1	.37	80.0	80.0	A04
20	8	10	1.57	.82	.86	-.2	.67	-.2	.42	80.0	80.0	A20
25	8	10	1.57	.82	.93	.0	.79	.0	.33	80.0	80.0	A25
34	8	10	1.57	.82	1.27	.7	1.49	.8	-.12	80.0	80.0	A34
63	8	10	1.57	.82	1.09	.3	.95	.2	.17	80.0	80.0	A63
65	8	10	1.57	.82	.91	-.1	.71	-.1	.37	80.0	80.0	A65
67	8	10	1.57	.82	.86	-.2	.67	-.2	.42	80.0	80.0	A67
70	8	10	1.57	.82	1.09	.4	1.08	.4	.13	80.0	80.0	A70
1	7	10	.99	.72	.90	-.3	.86	-.1	.39	80.0	70.1	A01
31	7	10	.99	.72	1.07	.3	1.02	.2	.21	80.0	70.1	A31
33	7	10	.99	.72	1.11	.5	1.05	.3	.17	60.0	70.1	A33
43	7	10	.99	.72	.74	-.9	.63	-.6	.58	80.0	70.1	A43
48	7	10	.99	.72	1.59	1.9	3.28	2.9	-.64	60.0	70.1	A48
60	7	10	.99	.72	1.06	.3	.99	.2	.23	80.0	70.1	A60
64	7	10	.99	.72	1.08	.4	1.07	.3	.19	60.0	70.1	A64
6	6	10	.50	.68	.72	-1.5	.66	-.9	.63	90.0	64.9	A06
7	6	10	.50	.68	.85	-.7	.78	-.5	.50	70.0	64.9	A07
10	6	10	.50	.68	1.03	.2	.96	.0	.30	50.0	64.9	A10
17	6	10	.50	.68	.86	-.7	.83	-.3	.47	90.0	64.9	A17
23	6	10	.50	.68	.72	-1.5	.66	-.9	.63	90.0	64.9	A23
37	6	10	.50	.68	1.32	1.5	1.28	.8	-.03	50.0	64.9	A37
39	6	10	.50	.68	.93	-.3	.86	-.2	.41	70.0	64.9	A39
66	6	10	.50	.68	.72	-1.5	.66	-.9	.63	90.0	64.9	A66
5	5	10	.04	.68	1.45	2.1	1.41	1.3	-.13	30.0	64.0	A04
8	5	10	.04	.68	1.61	2.7	2.02	2.7	-.41	30.0	64.0	A08
16	5	10	.04	.68	1.25	1.2	1.21	.8	.08	50.0	64.0	A16
21	5	10	.04	.68	.85	-.7	.82	-.5	.52	90.0	64.0	A21
24	5	10	.04	.68	.66	-1.9	.62	-1.4	.72	90.0	64.0	A24
30	5	10	.04	.68	1.03	.2	.98	.0	.33	70.0	64.0	A30
38	5	10	.04	.68	1.10	.5	1.06	.3	.25	50.0	64.0	A38
40	5	10	.04	.68	1.12	.6	1.06	.3	.24	50.0	64.0	A40
44	5	10	.04	.68	.66	-1.9	.62	-1.4	.72	90.0	64.0	A44
46	5	10	.04	.68	.93	-.3	.90	-.2	.43	70.0	64.0	A46
57	5	10	.04	.68	.88	-.5	.83	-.5	.49	70.0	64.0	A57
71	5	10	.04	.68	.86	-.6	.81	-.6	.51	70.0	64.0	A71
9	4	10	-.42	.69	1.16	.7	1.17	.6	.20	60.0	67.4	A09
11	4	10	-.42	.69	.95	-.1	.88	-.3	.44	60.0	67.4	A11
12	4	10	-.42	.69	.97	.0	1.00	.1	.40	60.0	67.4	A12
26	4	10	-.42	.69	.86	-.5	.80	-.6	.53	60.0	67.4	A26
36	4	10	-.42	.69	.89	-.3	.94	-.1	.47	80.0	67.4	A36
45	4	10	-.42	.69	.74	-1.0	.68	-1.0	.66	80.0	67.4	A45
47	4	10	-.42	.69	.90	-.3	.83	-.4	.50	80.0	67.4	A47
49	4	10	-.42	.69	1.13	.6	1.10	.4	.24	60.0	67.4	A49
59	4	10	-.42	.69	1.24	.9	1.25	.8	.11	60.0	67.4	A59
13	3	10	-.94	.75	.83	-.4	.74	-.5	.59	80.0	75.2	A13
15	3	10	-.94	.75	1.14	.5	.99	.1	.29	60.0	75.2	A15
32	3	10	-.94	.75	.83	-.4	.80	-.3	.57	80.0	75.2	A32
41	3	10	-.94	.75	.68	-.8	.58	-1.0	.74	80.0	75.2	A41
52	3	10	-.94	.75	1.06	.3	1.17	.5	.29	80.0	75.2	A52
56	3	10	-.94	.75	.66	-.9	.56	-1.0	.76	80.0	75.2	A56
62	3	10	-.94	.75	1.57	1.5	1.70	1.5	-.24	60.0	75.2	A62
14	2	10	-1.57	.86	.69	-.5	.55	-.6	.71	90.0	82.5	A14
22	2	10	-1.57	.86	.79	-.3	.76	-.2	.58	90.0	82.5	A22
42	2	10	-1.57	.86	1.62	1.2	1.99	1.4	-.32	70.0	82.5	A42
50	2	10	-1.57	.86	.79	-.3	.76	-.2	.58	90.0	82.5	A50
51	2	10	-1.57	.86	1.24	.6	.93	.1	.23	70.0	82.5	A51
54	2	10	-1.57	.86	1.45	.9	1.41	.8	-.06	70.0	82.5	A54
58	2	10	-1.57	.86	.63	-.6	.47	-.8	.78	90.0	82.5	A58
61	2	10	-1.57	.86	1.45	.9	1.88	1.3	-.14	70.0	82.5	A61
69	2	10	-1.57	.86	.63	-.6	.47	-.8	.78	90.0	82.5	A69
73	2	10	-1.57	.86	.79	-.3	.76	-.2	.58	90.0	82.5	A73
74	2	10	-1.57	.86	1.40	.9	1.21	.5	.03	70.0	82.5	A74
2	1	10	-2.51	1.12	.52	-.5	.22	-.6	.82	90.0	90.0	A02
18	1	10	-2.51	1.12	1.21	.5	.83	.2	.21	90.0	90.0	A18
27	1	10	-2.51	1.12	1.35	.7	1.52	.8	-.05	90.0	90.0	A27
28	1	10	-2.51	1.12	1.26	.6	1.01	.4	.13	90.0	90.0	A28
29	1	10	-2.51	1.12	1.39	.7	2.02	1.1	-.17	90.0	90.0	A29
35	1	10	-2.51	1.12	.52	-.5	.22	-.6	.82	90.0	90.0	A35
55	1	10	-2.51	1.12	.52	-.5	.22	-.6	.82	90.0	90.0	A55
68	1	10	-2.51	1.12	1.28	.6	1.08	.5	.10	90.0	90.0	A68
72	1	10	-2.51	1.12	1.39	.7	2.02	1.1	-.17	90.0	90.0	A72
53	0	10	-3.90	1.90	MINIMUM ESTIMATED MEASURE							A53
75	0	10	-3.90	1.90	MINIMUM ESTIMATED MEASURE							A75
MEAN	4.3	10.0	-.40	.82	1.01	.1	.99	.1		75.2	74.5	
S.D.	2.4	.0	1.40	.23	.28	.9	.48	.8		14.6	9.1	

Art. Revista Investigacion Educativa N 24.indd 145 24/05/2010 03:30:08 p.m.

Tabla IV

```
GUTTMAN SCALOGRAM OF RESPONSES:
alumno |pregunta
       |         1
       |3528471906
       |----------
    19 +1111101111  A19
     3 +1111111010  A03
     4 +1111110101  A04
    20 +1111110110  A20
    25 +1111011110  A25
    34 +1011101111  A34
    63 +1111010111  A63
    65 +1111110101  A65
    67 +1111110110  A67
    70 +1110111011  A70
     1 +1101111100  A01
    31 +1101101110  A31
    33 +1101101101  A33
    43 +1111110100  A43
    48 +0011011111  A48
    60 +1110011110  A60
    64 +1011110101  A64
```

Tabla V

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	pregunta
6	20	73	1.01	.30	1.06	.4	.94	-.1	.42	75.3	76.8	P6
10	22	73	.83	.29	.96	-.3	.81	-.6	.50	76.7	75.8	P10
9	24	73	.67	.28	.80	-1.7	.70	-1.2	.60	84.9	74.8	P9
1	27	73	.43	.28	1.25	1.9	1.57	2.3	.31	69.9	73.4	P1
4	31	73	.13	.27	1.10	.8	1.06	.4	.46	72.6	72.2	P4
7	31	73	.13	.27	.89	-.9	.98	-.1	.56	75.3	72.2	P7
8	36	73	-.25	.27	.98	-.2	.98	-.1	.54	74.0	72.3	P8
2	37	73	-.32	.27	.96	-.3	.92	-.4	.55	72.6	72.4	P2
5	40	73	-.54	.27	1.08	.7	1.07	.5	.49	67.1	73.0	P5
3	58	73	-2.09	.33	.89	-.6	.86	-.3	.56	83.6	82.0	P3
MEAN	32.6	73.0	.00	.28	1.00	.0	.99	.0		75.2	74.5	
S.D.	10.5	.0	.85	.02	.12	1.0	.22	.9		5.3	2.9	

Art. Revista Investigacion Educativa N 24.indd 146 24/05/2010 03:30:08 p.m.

Tabla VI

```
MOST MISFITTING RESPONSE STRINGS
pregunta   OUTMNSQ  |alumno
                    |1763264 31 6577655427622 21
                    |9034548862922431402228978
                  high----------------------------
     1 P1      1.57 A|..............1.1.1.1.1...
     4 P4      1.06 B|..0.0........11..1.1......
     5 P5      1.07 C|...0.00..................1
     6 P6       .94 D|........111....1..........
     7 P7       .98 E|0..0.....................1.
     8 P8       .98 e|.0...................1....
     2 P2       .92 d|.......................1..
    10 P10      .81 c|...........11.............
     3 P3       .86 b|......00..................
     9 P9       .70 a|..................1.......
                    |----------------------low-
                    |1763264831965776554276222 1
                    |9034548  62  224314022289878
```

B_{72}=-2.51 y B_{29}=-2.51. Como es fácil notar los niveles de habilidad de estos cuatro alumnos están muy por debajo del nivel de dificultad de P1. El análisis de las preguntas de la prueba evidenció contenidos de aritmética. Los mismos que fueron trabajados durante el ciclo académico donde se aplicó la prueba. Si bien encontramos distintos temas como regla de tres, conjuntos, porcentajes y proporciones, entre otros, todos ellos corresponden a la misma área. No podemos decir que P1 mida algo distinto que el resto de las preguntas. En el modelo de Rasch las medidas se encuentran a lo largo de una dimensión empírica definida por un consenso de los datos. En nuestro análisis se obtuvo que la varianza empírica explicada por las medidas resultó del 51.2% que es casi igual al 50.6%, cantidad de varianza explicada si los datos se ajustarían perfectamente al modelo.

Otra de las ventajas del modelo de Rasch es que las medidas de los alumnos y preguntas, al estar en las mismas unidades, se pueden ubicar en la misma escala. El mapa mostrado en la figura I constituye una forma de analizar los resultados de la prueba gráficamente. A cada lado de la línea punteada central se encuentran los alumnos (izquierda)

Art. Revista Investigacion Educativa N 24.indd 147 24/05/2010 03:30:08 p.m.

y las preguntas (derecha) ordenados según sus medidas estimadas. En este caso se muestran valores comprendidos entre -3 y 3 logits, los alumnos con mayor nivel de habilidad y las preguntas con mayor nivel de dificultad se encuentran en la parte superior. Los alumnos de menor habilidad y las preguntas más fáciles son ubicados en la parte inferior. Esta distribución ordenada nos permite hacer una interpretación de la prueba al comparar los niveles de habilidad con las dificultades de la preguntas. En el extremo superior de la escala, por encima de los 1.5 logits, encontramos 10 alumnos y en el extremo inferior, por debajo de los -2.5 logits, encontramos 11 alumnos. Para este grupo no se encontró ninguna pregunta cercana a sus niveles de habilidad. Las preguntas estuvieron concentradas entre los -0.5 y 1.0 logits. Dos preguntas (P4 y P7) comparten la misma ubicación y la pregunta P3 se ubicó bastante alejada de las demás. Estas observaciones permiten hacer mejoras en la prueba en aplicaciones futuras. Las preguntas deberían cubrir un mayor intervalo y procurar la misma separación entre ellas.

Figura I

Art. Revista Investigacion Educativa N 24.indd 148 24/05/2010 03:30:08 p.m.

3.2. Comparación de las medidas estimadas y los puntajes

Como se señaló anteriormente alumnos con el mismo puntaje presentan la misma medida estimada, el coeficiente de correlación de Pearson entre ambos fue calculado en 0.987 indicando una fuerte correlación positiva. Mientras que los puntajes forman una escala ordinal las medidas estimadas corresponden a medidas lineales formando una escala de intervalo. A partir de la información mostrada en la tabla III se ha construido la tabla VII. De acuerdo con estos resultados a una diferencia de 1 punto entre puntajes no le corresponde una diferencia constante entre sus correspondientes medidas estimadas. Esto pone en evidencia que los puntajes brutos forman una escala ordinal y que si bien existen diferencias de grado en el rendimiento de alumnos con distintos puntajes, estas diferencias no reflejan cuanto más dominio de aritmética ellos poseen. Al contrario que los puntajes, las medidas estimadas corresponden a medidas lineales que forman una escala de intervalo.

Tabla VII

Alumno	Score	Measure	Diferencia score	Diferencia measure
A19	9	2.42	1	0.85
A03	8	1.57		
A70	8	1.57	1	0.58
A01	7	0.99		
A64	7	0.99	1	0.49
A06	6	0.50		
A66	6	0.50	1	0.46
A05	5	0.04		
A71	5	0.04	1	0.46
A09	4	-0.42		
A59	4	-0.42	1	0.52
A13	3	-0.94		
A62	3	-0.94	1	0.63
A14	2	-1.57		
A74	2	-1.57	1	0.94
A02	1	-2.51		
A72	1	-2.51	1	1.39
A53	0	-3.90		

Cuanto mayor sea el nivel de dominio del alumno, mayor será la probabilidad de responder correctamente las preguntas y por tanto de tener una mayor puntuación en la prueba. Existe una relación entre la

Art. Revista Investigacion Educativa N 24.indd 149 24/05/2010 03:30:09 p.m.

medida estimada del alumno y la probabilidad de acierto a las preguntas. La tabla VIII muestra las probabilidades - calculadas con la fórmula I – de respuesta correcta de los alumnos A70 (B70=1.57), A05 (B5=0.04) y A13 (B13=-0.94) frente a las preguntas P4 (D4=0.13) y P2 (D2=-0.32).

Tabla VIII

Alumno	Probabilidad de responder correctamente	
	P4	P2
A70	0.808	0.869
A05	0.478	0.589
A13	0.255	0.350

El logaritmo natural del ratio de la probabilidad de responder correctamente una pregunta a la probabilidad de no responderla correctamente proporciona una estimación del dominio del alumno de acuerdo con la expresión:

$$D_{si} = 1n\left(\frac{P_{si}}{1 - P_{si}}\right)$$

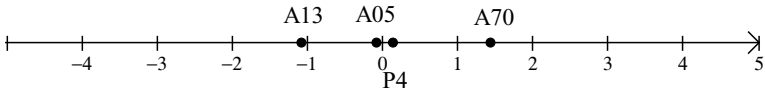
Utilizando esta última expresión y con los datos de la tabla VIII se han estimado los niveles de dominio de los alumnos en los temas tratados en la prueba. Los resultados se muestran en la tabla IX.

Tabla VIII

Alumno	Con referencia a	
	P4	P2
A70	1.437	1.892
A05	-0.088	0.360
A13	-1.072	-0.619

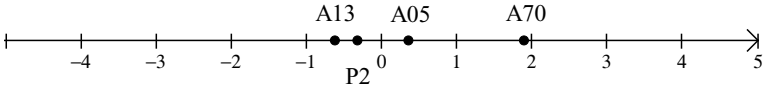
Las figuras 2 y 3 muestran la ubicación, en una misma escala, de los niveles de dominio estimados para cada alumno y el nivel de dificultad de la pregunta con la que fueron estimados.

Figura 2



Art. Revista Investigacion Educativa N 24.indd 150 24/05/2010 03:30:09 p.m.

Figura 3



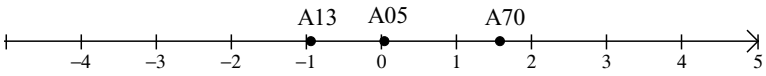
De acuerdo con los resultados de la tabla VIII, los niveles de dominios estimados de los alumnos no son iguales; sin embargo, la comparación de las figuras 2 y 3 evidencia que sus posiciones relativas en la escala son las mismas. En ambos casos A70 se encuentra a la derecha de A05 y éste a la derecha de A13. Una observación más profunda nos haría ver que la distancia entre A70 y A05 es la misma en ambas escalas y está dada por la diferencia entre sus niveles de dominio. En el primer caso esta distancia es $Dist_{(A70)(A05)} = 1.437 - (-0.088) = 1.525$ y en el segundo caso será $Dist_{(A70)(A05)} = 1.892 - 0.360 = 1.532$ las cuales son aproximadamente iguales. Al comparar las distancias dos a dos de los tres alumnos en los dos casos anteriores podemos observar que ellas se mantienen prácticamente constantes. La tabla IX muestra las diferencias entre los niveles de dominio estimados en cada caso.

Tabla IX

Distancia entre	Con referencia a	
	P4	P2
A70 y A05	1.525	1.532
A05 y A13	0.984	0.979
A70 y A13	2.509	2.511

Si bien existen pequeñas diferencias entre los valores de las columnas estas son debidas a la aproximación en los cálculos previos. De acuerdo a lo anterior podemos desprender que la estimación del nivel de dominio de los alumnos es independiente de la pregunta con la cual fue estimada. La ubicación de los alumnos y sus distancias relativas se mantienen si lo hacemos con referencia al total de preguntas en la prueba tal como se muestra en la figura 4.

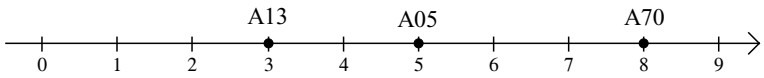
Figura 4



Art. Revista Investigacion Educativa N 24.indd 151 24/05/2010 03:30:09 p.m.

En la figura 5 se muestran las ubicaciones de los alumnos A70, A05 y A13 en la escala ordinal con referencia al puntaje bruto en la prueba. Nótese que si bien las posiciones se mantienen sus diferencias (distancias) son mayores que las encontradas anteriormente. El alumno A70 parece dominar el tema más que A05 y éste a su vez más que A13.

Figura 5



La revisión del escalograma nos permite observar las respuestas dadas por los alumnos a las preguntas P4 y P2. La información se resume en la tabla X.

Tabla X

Alumno	P4	P2
A70	1	1
A05	0	0
A13	0	1

Nótese ahora que las ubicaciones de los alumnos cambian si se hace con referencia a la P4 (figura 6), P2 (figura 7) y ellas juntas (figura 8).

Figura 6

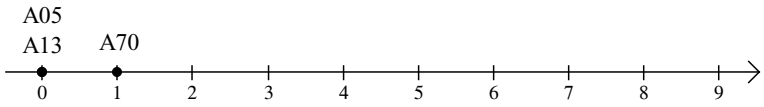
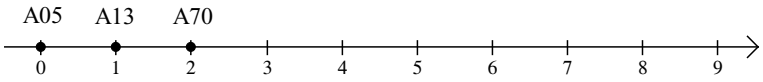


Figura 7

Art. Revista Investigacion Educativa N 24.indd 152 24/05/2010 03:30:09 p.m.

Figura 8



En la figura 6, la ubicación de A13 y A05 parece decirnos que ellos dominan por igual el tema de la pregunta. En la figura7, parece decirnos que A70 y A13 tienen igual dominio y que A13 domina más que A05 el tema de la pregunta. Luego, en la figura 8 la ubicación de A70 indica un mayor dominio que A13 y éste más que A05, además que las distancias entre ellos son iguales. Nótese que la información obtenida con la medida ordinal es imprecisa e incluso contradictoria tal como se observa al comparar las figuras 5 y 8. En estos casos los niveles de dominio de los alumnos depende del instrumento con el que fue estimado.

Los puntajes obtenidos por los alumnos se pueden reportar en una escala de 0 a 20. Dado que la prueba consta de 10 preguntas la relación nota1=2(score) permite transformar los scores en la acostumbrada nota vigesimal. Sin embargo, tal como se había comentado, los puntajes brutos reportan solo una medida ordinal y esta no dice nada acerca de las diferencias entre los puntajes de los alumnos. Las medidas estimadas con el modelo de Rasch constituyen una medida de intervalo y en una escala de este tipo la diferencia de una unidad significa lo mismo a lo largo de toda la escala. Una limitación que presenta el modelo de Rasch es que las medidas estimadas son números reales y por lo general estamos acostumbrados a reportar y leer puntajes en una escala vigesimal. Esta limitación se puede salvar con una transformación lineal de las medidas estimadas a la nota vigesimal. La nota 0 corresponde al score 0, en este caso a una medida estimada en -3.90; la nota 20 correspondería al score 10, el cual con ayuda del ministep fue estimado en 3.73. Si representamos "nota 2" la nota vigesimal y con "measure" la medida estimada, a partir de los datos presentados obtenemos la relación nota 2 = 2.6 (measure) + 10.2 que permite transformar las medidas estimadas por el modelo de Rasch a notas vigesimales. La tabla XI muestra los resultados de dichas transformaciones.

Art. Revista Investigacion Educativa N 24.indd 153 24/05/2010 03:30:10 p.m.

Tabla XI

Score	Measure	nota 1	nota 2
9	2.42	18	17
8	1.57	16	14
7	0.99	14	13
6	0.50	12	12
5	0.04	10	10
4	-0.42	8	9
3	-0.94	6	8
2	-1.57	4	6
1	-2.51	2	4
0	-3.90	0	0

Como se puede apreciar, las notas (nota 1) obtenidas a partir de los puntajes brutos (scores) no son las mismas que las notas (nota 2) obtenidas a partir de las medidas estimadas (measure). Así por ejemplo un alumno con score 8 sería calificado de forma tradicional con una nota de 16, sin embargo la nota obtenida a partir de su medida estimada sería de 14 puntos. De acuerdo con lo expuesto en el presente articulo este segundo resultado refleja mejor el dominio del alumno en los contenidos tratados en la prueba analizada.

4. CUESTIÓN FINAL

La evaluación objetiva de los aprendizajes supone su medición. La información recogida debe ser analizada cuidadosamente. Si bien no debemos reducir la evaluación a la aplicación de prácticas y exámenes, mucho menos debemos considerar que el número asignado a su calificación es un indicador absoluto del logro del alumno. Los puntajes brutos de los alumnos obtenidos en las pruebas constituyen una medida ordinal y esta nos dice muy poco acerca de los niveles de dominio de los alumnos del tema tratado en la prueba. De acuerdo a esto se debe tener cuidado al emitir juicios acerca de los aprendizajes de los alumnos a partir de estos puntajes. Usar el modelo de Rasch y generar medidas lineales nos proporciona información adicional que hace posible una medición más objetiva.

Art. Revista Investigacion Educativa N 24.indd 154 24/05/2010 03:30:10 p.m.

BIBLIOGRAFÍA

Andrich, D. (1988). Rasch models for measurement. Sage University Paper series on Quantitative Applications in the Social Sciences, series N° 07-068, California, Sage Publications, pp. 20-21.

Bond, T. y Fox, C. (2007). Applying the Rasch Model. Lawrence Eribaum Associates, Publishers, New Jersey, p. 41.

Frisbie, D. (1988). Reliability of scores from teacher-made tests. Educational Measurement: Issues and Practice, National Council on Measurement in Education, 7(1), pp. 25-35.

Glass, G. y Stanley, J. (1970). Statistical methods in education and psychology. Prentice Hall. New Jersey, p. 10.

Hambleton, Swaminathan y Rogers (1991). Fundamentals of Item Response Theory. Sage Publications, California, pp. 2-3.

Muñiz, J. (1997). Introducción a la Teoría de Respuesta a los Ítems. Ediciones Pirámide, Madrid, p. 17.

Sears, F. y Zemansky, M. (1981). Física General, 5ta edición. Aguilar SA Ediciones, Madrid, p. 3.

Tristán, A. (2001). Análisis de Rasch para Todos. CENEVAL, México, p. 11.

Wright, B. y Mok, M. (2004). An Overview of the Family of Rasch Measurement Models. En Introduction to Rasch Measurement, Smith, E. y Smith, R. (Ed), JAM Press, pp. 2-24.

Wright, B. y Stone, M. (1999). Measurement Essentials. WIDE RANGE, Inc., Wilmington, p. 9.

Art. Revista Investigacion Educativa N 24.indd 155 24/05/2010 03:30:10 p.m.